

МАСШТАБИРУЕМЫЙ АЛГОРИТМ МЕТРОПОЛИСА-ХАСТИНГСА

Находнов Максим Сергеевич

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: nakhodnov17@gmail.com

Научный руководитель — Кропотов Дмитрий Александрович

Множество современных алгоритмов машинного обучения, как с учителем, так и без учителя, основывается на методах Байесовского вывода. Однако, такие модели зачастую невозможно аналитически вычислить и, как следствие, они требуют использования приближённых методов. Так, например, Методы Монте-Карло по схеме марковских цепей (МСМС) используются для получения выборки из целевого апостериорного распределения. Примером такого метода является алгоритм Метрополиса-Хастингса. Однако, такой метод вычислительно неэффективен, так как для выполнения одной итерации алгоритма требуется полный проход по всем данным. Учитывая, что МСМС методы требуют значительного числа итераций для «разогрева», а затем для получения некоррелированных точек, такой подход может оказаться неприменимым на практике.

Трудоёмкость одной итерации алгоритма Метрополиса-Хастингса зависит, с одной стороны, от сложности семплирования из распределения перехода, а с другой — от трудоёмкости вычисления вероятности новой точки при условии предыдущей. Для применения схемы Метрополиса-Хастингса к большим объёмам данных каждое из этих двух действий должно выполняться эффективно, не требуя прохода по всей обучающей выборке.

В статье [2] была предложена модификация схемы Метрополиса-Хастингса — Minibatch Metropolis Hastings, основанная на тесте Баркера, которая позволяет выполнять семплирование без просмотра всей выборки.

В данной работе предлагается метод, который расширяет схему Minibatch Metropolis Hastings за счёт расширения класса acceptance function с помощью семейства функций, удовлетворяющих лемме Баркера. За счёт нового гиперпараметра, который определяет функцию из предложенного семейства, появляется возможность подстраивать работу схемы Метрополиса-Хастингса под конкретную задачу, меняя баланс между скоррелированностью семплов и скоростью их генерации.

Для оценки качества работы алгоритмов семплирования используется метрика ESS [3]. Эта величина позволяет оценить скоррелированность точек в марковской цепи, а именно: она отражает число нескоррелированных семплов из данного распределения, которые были бы эквивалентны набору скоррелированных точек из данной цепи.

Предложенное расширение алгоритма позволяет ускорить оригинальный Minibatch Metropolis Hastings без уменьшения ESS, а в некоторых случаях и уменьшить величину корреляции между точками марковской цепи.

Литература

1. Eddington A.S. On a Formula for Correcting Statistics for the Effects of a known Probable Error of Observation // Monthly Notices of the Royal Astronomical Society. — 1913. — 03. — Vol. 73, no. 5. — Pp. 359–360.
2. Seita Daniel, Pan Xinlei, Chen Haoyu, Canny John. An efficient minibatch acceptance test for metropolis-hastings. — 2016.
3. Thompson Madeleine B. A comparison of methods for computing autocorrelation time. — 2010.